MST121 Chapter D3

The Open University

A first level interdisciplinary course

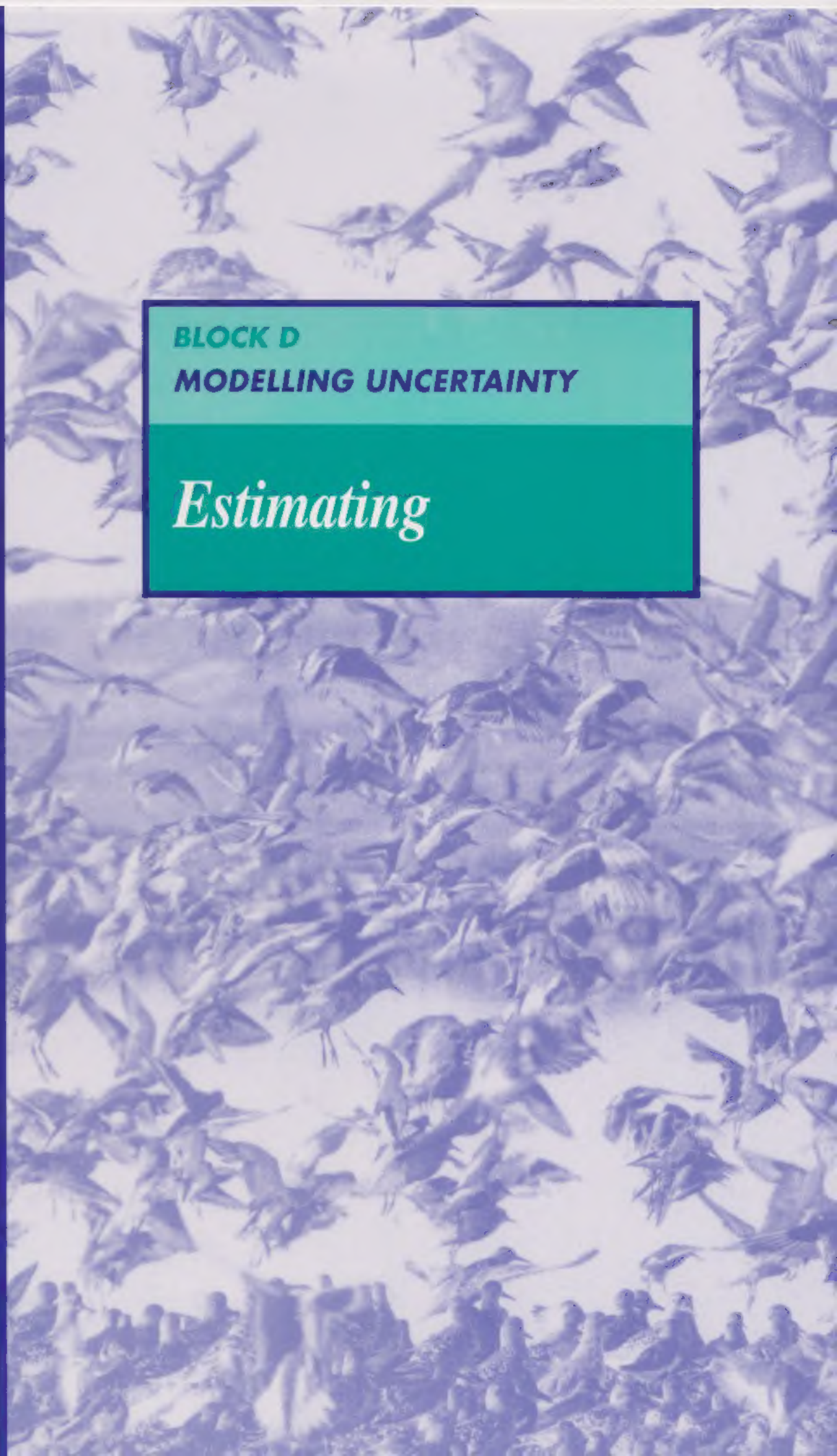
Using Mathematics

CHAPTER D3

BLOCK D

MODELLING UNCERTAINTY

Estimating



The Open University

A first level interdisciplinary course

Using Mathematics

BLOCK D

MODELLING UNCERTAINTY

Estimating

CHAPTER D3

Prepared by the course team

About this course

This course, MST121 *Using Mathematics*, and the courses MU120 *Open Mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. Further details may be obtained from the address below.

MST121 uses the software program Mathcad (MathSoft, Inc.) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of the course, and its use is covered in the associated Computer Book.

The Open University, Walton Hall, Milton Keynes MK7 6AA.

First published 1997. Reprinted 1997, 1998, 1999, 2000, 2001.

Edited, designed and typeset by The Open University using the Open University TEX System.

Printed in the United Kingdom by The Burlington Press, Foxton, Cambridge CB2 6SW.

ISBN 0 7492 7890 0

This text forms part of an Open University First Level Course. If you would like a copy of *Studying with The Open University*, please write to the Course Enquiries Data Service, PO Box 625, Dane Road, Milton Keynes MK1 1TY. If you have not already enrolled on the Course and would like to buy this or other Open University material, please write to Open University Educational Enterprises Ltd, 12 Cofferidge Close, Stony Stratford, Milton Keynes MK11 1BY, United Kingdom.

1.6

Contents

Study guide

You should schedule three study sessions for your work on this chapter, of which the first will include studying a video band and the third will use the computer. The study pattern which we recommend is as follows.

Study session 1: Section 1. You will need access to a video player for this session.

Study session 2: Section 2.

Study session 3: Sections 3 and 4. You will need access to your computer, together with the statistics software and Computer Book D for Section 3.

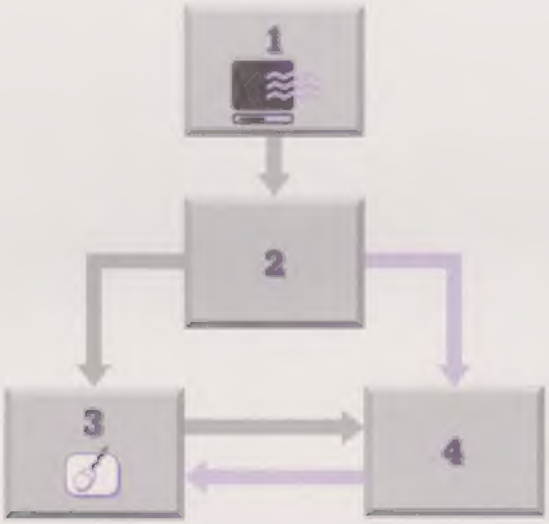
The television programme *The Spiral of Silence* is related to Section 4.

Section 4 is very short and does not require the computer. You may, if you wish, study it immediately after Section 2. An alternative study pattern is thus as follows.

Alternative study session 1: Section 1 (video).

Alternative study session 2: Sections 2 and 4.

Alternative study session 3: Section 3 (computer work).



Introduction

Since 1937, when British Gallup was founded in the UK, opinion polls on all sorts of topics have been carried out – on divorce, on mercy killings, on compulsory military training and, indeed, on almost any political issue of current public interest you could name. However, when opinion polls are mentioned, most people immediately think of the polls on voting intentions which are held during election campaigns. In fact, such polls are held regularly between elections too, but it is the polls held at election times that receive the most publicity and that are remembered most if their predictions turn out to be wrong. In the UK in 1992, for instance, four eve-of-election polls underestimated the Conservative party's share of the vote by an average of 4%, and overestimated the Labour party's share by a similar amount. Instead of the 0.8% Labour lead over the Conservatives that had been forecast, the election produced a substantial Conservative lead over Labour of 7.6%. As a result, the Conservatives were unexpectedly returned to power.

It would not have been surprising if the shares of votes predicted by the polls had differed a little from the actual result – poll predictions are based on interviews with a sample of electors, and there is no guarantee that the voting intentions of the sample will reflect precisely the voting behaviour of those who actually vote. Indeed, in no general election from 1945 onwards have the polls predicted the political parties' shares of the vote with absolute accuracy. Nor do the polling organisations claim such accuracy for their predictions: they always attach a 'margin of error' to their poll results (although this is not always reported by the media). This 'margin of error' provides an indication of the reliability of the predictions and of how much confidence can be placed in them. The 'margin of error' depends on the sample size – the larger the sample, the smaller the 'margin of error'. So, intuitively, the more people that are interviewed, the more confidence we can place in a poll's predictions.

In 1992, the BBC calculated a 'poll of polls' by combining the results of four eve-of-election polls (conducted by Gallup, MORI, NOP and ICM). The results of the poll of polls and the election itself are shown below. The predictions were based on the results of interviews with 8141 people.

	Conservative	Labour	Liberal Democrat
Poll of polls	38%	39%	19%
Election result	43%	35%	18%

The 'margin of error' quoted was ±1.5%. This means that a share of the vote which differed from the predicted share by as much as 1.5 percentage points up or down would not be regarded as surprising. So the poll suggested that the Conservative share, for instance, might turn out to be anywhere between 36.5% and 39.5%. However, the actual result was well outside this range at 43%! Why were the polls so far out in their predictions? The television programme *The Spiral of Silence* looks at a number of possible explanations. And in this chapter, we shall discuss the idea behind a 'margin of error'.

The purpose of a poll is to estimate the voting shares of the various political parties. A sample of electors is interviewed, and the responses obtained are used to estimate the voting shares of the parties; the quality of these estimates is indicated by attaching a 'margin of error', thus providing a range of plausible values for the parties' voting shares. The

design of a political opinion poll is sophisticated, and the calculation of results is correspondingly complicated. However, the principle of collecting a sample of data and using it to calculate a range of plausible values for some unknown quantity is a general one. In this chapter, we consider the simpler situation of estimating the population mean, given a random sample from a population.

In Chapter D2, you saw that the sample mean can be used to estimate the value of an unknown population mean. However, it is always possible that a particular sample is unrepresentative of the population from which it is drawn, so the sample mean may not be a good estimate of the population mean. So how inaccurate might it be? In this chapter, we look at how to use a sample of data to obtain a range of plausible values for the population mean – this range of values is called a *confidence interval for the population mean*.

Suppose that a sample is drawn from a population, and the sample mean is calculated: this sample mean is our estimate of the unknown population mean. For instance, in Chapter D2, the mean height of a sample of 1000 Cambridge men was used to estimate the mean height of the population of all Cambridge men in 1902. But how good is this estimate? Is it likely to be close to the true value of the population mean, or might it differ considerably from the population mean?

If the heights of a second sample of 1000 Cambridge men had been obtained, then the mean height of this sample would very probably have been different from the mean height of the first sample: the sample mean will vary from sample to sample. But by how much does it vary? Are all sample means close together, so that we can be confident that the mean height of any sample (including ours) of 1000 men is close to the mean height of all Cambridge men – the population mean? Or does the sample mean vary greatly from sample to sample? To answer the question 'How good is the estimate?', we need to know how sample means vary. This is investigated in Section 1: your work in this section includes studying a video band.

The results obtained in Section 1 are applied in Section 2 to develop a method for obtaining a range of plausible values for an unknown population mean – that is, a confidence interval for the population mean. We shall then be able to assess how 'good' the mean height of the 1000 Cambridge men is as an estimate of the mean height of all Cambridge men in 1902, by calculating a confidence interval for this population mean. In Section 3, you will learn how to use OUStats to obtain confidence intervals. Finally, in Section 4, we return to opinion polls to discuss very briefly what is meant by a 'margin of error' in that context, and start to explore why the opinion polls predicted the wrong result in the 1992 general election.

The learning file theme for this chapter and the next is 'distinguishing different ideas with similar names'; one way you can think about this and demonstrate your ideas is by designing a short report for the theme. A main aim for such an activity is to help you to become familiar with and understand the fact that some ideas with similar names are in fact different. When trying to explain these ideas to other people, especially those who are unfamiliar with aspects of statistics, you need to be careful about the language you use – that is, the technical terms – so that appropriate, correct and precise meanings are provided. Writing a short report can be helpful to you by confirming and checking your own understanding.

You came across an example of distinguishing different ideas with similar names in Chapter D2. There, we distinguished between the mean of a population and the mean of a sample, and between the standard deviation of a population and the standard deviation of a sample. In Chapters D3 and D4, you will meet the mean and standard deviation of distributions which are linked to but different from the populations from which samples are drawn.

To understand the main statistical techniques that are introduced in these two chapters, you need to be able to explain what the different means and standard deviations are, and how they are related. Thinking about how to present particular ideas to other people, in writing and using diagrams, can help you to reflect on the different uses of terms such as 'standard deviation', and how to distinguish them.

1 Populations and samples

To study this section, you will need access to a video player together with Video Band D.

In practice, the actual distribution of a population is usually unknown indeed, the purpose of taking a sample may well be to obtain information about the population as a whole. In this section, in order to gain some insight into how sample means are related to the population from which the samples were drawn, we shall suppose that the distribution of a population is known. We shall investigate patterns in the means of samples from the population; and we shall consider various sample sizes. The relationships observed between a population and the means of samples from the population will be used in Section 2 to make inferences about a population from a sample of data.

1.1 All possible samples

Example 1.1 How many cars?

Suppose that a large motoring organisation wants to estimate the mean number of cars per household for its members. It plans to ask a sample of members how many cars there are in their households, and to use the sample mean as an estimate of the mean number of cars per household for all its members.

Begin by considering a very simple situation. Suppose that, in fact, half the members' households have one car and the other half have two cars. So the mean number of cars per household is 1.5. The distribution of the number of cars per household – the population distribution – is illustrated in Figure 1.1.

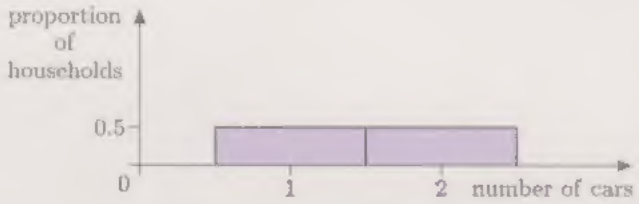


Figure 1.1 The population distribution

Now suppose that a sample of size 2 is drawn from this population. There are four possible outcomes; these are listed in the first two columns of Table 1.1. For now, we assume that each of the four different possible samples is equally likely (we discuss this assumption further later); so if all possible samples of size 2 were drawn from this population, then one quarter of them would be of each type. Also included in Table 1.1 is the mean of each sample and, in the final column, the proportion of samples which are of each type.

Table 1.1 Possible samples of size 2

Possible samples			
First household	Second household	Sample mean	Proportion of samples with this mean
1	1	1	$\frac{1}{4}$
1	2	1.5	$\frac{1}{4}$
2	1	1.5	$\frac{1}{4}$
2	2	2	$\frac{1}{4}$

Table 1.2 contains the possible values of the sample mean and, for each value, the proportion of samples which have this mean. This gives the distribution of the sample mean for samples of size 2 from this population. This distribution is called the **sampling distribution of the mean** for samples of size 2.

Table 1.2 Sample means for samples of size 2

Sample mean	Proportion of samples with this mean
	$\frac{1}{4}$
1.5	$\frac{1}{2}$
2	$\frac{1}{4}$

A picture of this distribution is shown in Figure 1.2.

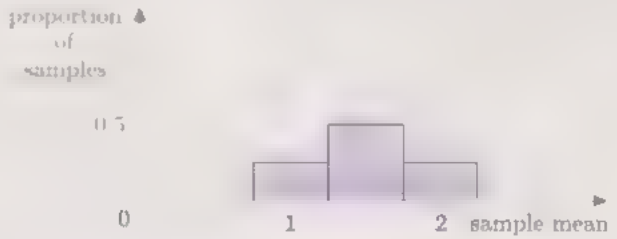


Figure 1.2 The sampling distribution of the mean for samples of size 2

Suppose that there are N possible different samples of size 2 altogether. Then $\frac{1}{4}N$ of these samples have mean 1, $\frac{1}{2}N$ have mean 1.5 and $\frac{1}{4}N$ have mean 2. So the average value (that is, the mean) of all the sample means is

$$\frac{1}{N}\sum xf = \frac{1}{N}\left(1 \times \tfrac{1}{4}N + 1.5 \times \tfrac{1}{2}N + 2 \times \tfrac{1}{4}N\right) = 1.5.$$

Notice that N cancelled out: we did not need to know how many possible samples of size 2 there were.

about 1.5, we can see intuitively that the mean is 1.5.

So the average – that is, the mean value of the sample mean for samples of size 2 is 1.5, the same as the mean of the population.

In order to obtain the results just given for the sampling distribution of the mean, an assumption about the population was made implicitly. This assumption is that the population is a large one, and is much larger than the size of the samples drawn from it. Essentially, we assumed that the population of households is so large that taking one household from the population does not affect the proportions of the remaining households with 1 car and 2 cars. We assumed that the probabilities of the possible outcomes of selecting a second household are the same as for the first household.

To see why this assumption is necessary, consider the following two situations.

First, suppose that the population is a small one consisting of only 20 households, 10 of which have 1 car and 10 of which have 2 cars. The probability that the first household selected has 1 car is $\frac{1}{2}$. If the first household selected has 1 car, then 9 of the remaining households have 1 car and 10 have 2 cars, so the probability that the second household selected has 1 car is $\frac{9}{19} \simeq 0.474$, not $\frac{1}{2}$. When the population is small, the probabilities of the different outcomes change noticeably as households are removed. And hence the probability that both households in a sample of size 2 have 1 car is $\frac{1}{2} \times 0.474 = 0.237$, not $\frac{1}{4} = 0.25$.

Now suppose that the population is much larger, say 200 000 households, and that half of them have 1 car and half have 2 cars. Again, the probability that the first household selected has 1 car is $\frac{1}{2}$. However, if the first household has 1 car, then 99 999 of the remaining 199 999 households have 1 car, so the probability that the second household also has 1 car is

$$\frac{99\,999}{199\,999} \simeq 0.499\,997\,5$$

which is very close to $\frac{1}{2}$. So for this large population, the error introduced by assuming that the probability that the second household selected has 1 car is $\frac{1}{2}$ is negligible. And hence the probability that both households in a sample of size 2 have 1 car is very close to $\frac{1}{4}$.

All the results in this section depend on the assumption that the population is much larger than the samples drawn from it.

Activity 1.1 Samples of size 4

This is similar to listing all the possible outcomes of tossing 4 coins or all the possible patterns of boys and girls in families of size 4. See Activity 3.5 of Chapter D1.

Complete the following table for samples of size 4 from the population in Figure 1.1. List all the possible different samples in the first column (there are 16 of them), then calculate the mean of each sample and enter it in the second column.

Sample values	Sample mean	Proportion of samples with this mean
1 1 1 1	1	$\frac{1}{16}$
1 1 1 2	1.25	$\frac{1}{16}$

Comment

The solution is given on page 37.

Table 1.3 shows the possible values of the sample mean for samples of size 4 together with, for each value, the proportion of samples which have this mean. The table gives the *sampling distribution of the mean* for samples of size 4 from the population in Figure 1.1. A picture of this distribution is shown in Figure 1.3.

Table 1.3 Sample means for samples of size 4

Sample mean	Proportion of samples
1	$\frac{1}{16}$
1.25	$\frac{4}{16}$
1.5	$\frac{6}{16}$
1.75	$\frac{4}{16}$
2	$\frac{1}{16}$

Figure 1.3 The sampling distribution of the mean for samples of size 4

This sampling distribution is also symmetrical about the value 1.5, so intuitively you can perhaps see that its mean is 1.5. Again, the mean of the sampling distribution is equal to the population mean: the average value of the sample mean for samples of size 4 is equal to the population mean.

If you are interested, you could check that the mean of this sampling distribution is 1.5, either by using the approach used for the sampling distribution for samples of size 2, or by using the formula for the mean of a distribution which was discussed briefly in Chapter D1 (formula (4.4)) and Chapter D2 (Subsection 2.1). However, in this course *you will not be required to calculate the mean of a probability distribution for yourself.* In general, when the mean of a distribution is needed, we shall simply state the required result.

Figure 1.4 contains diagrams of the sampling distributions of the mean for samples of sizes 9, 25 and 100.

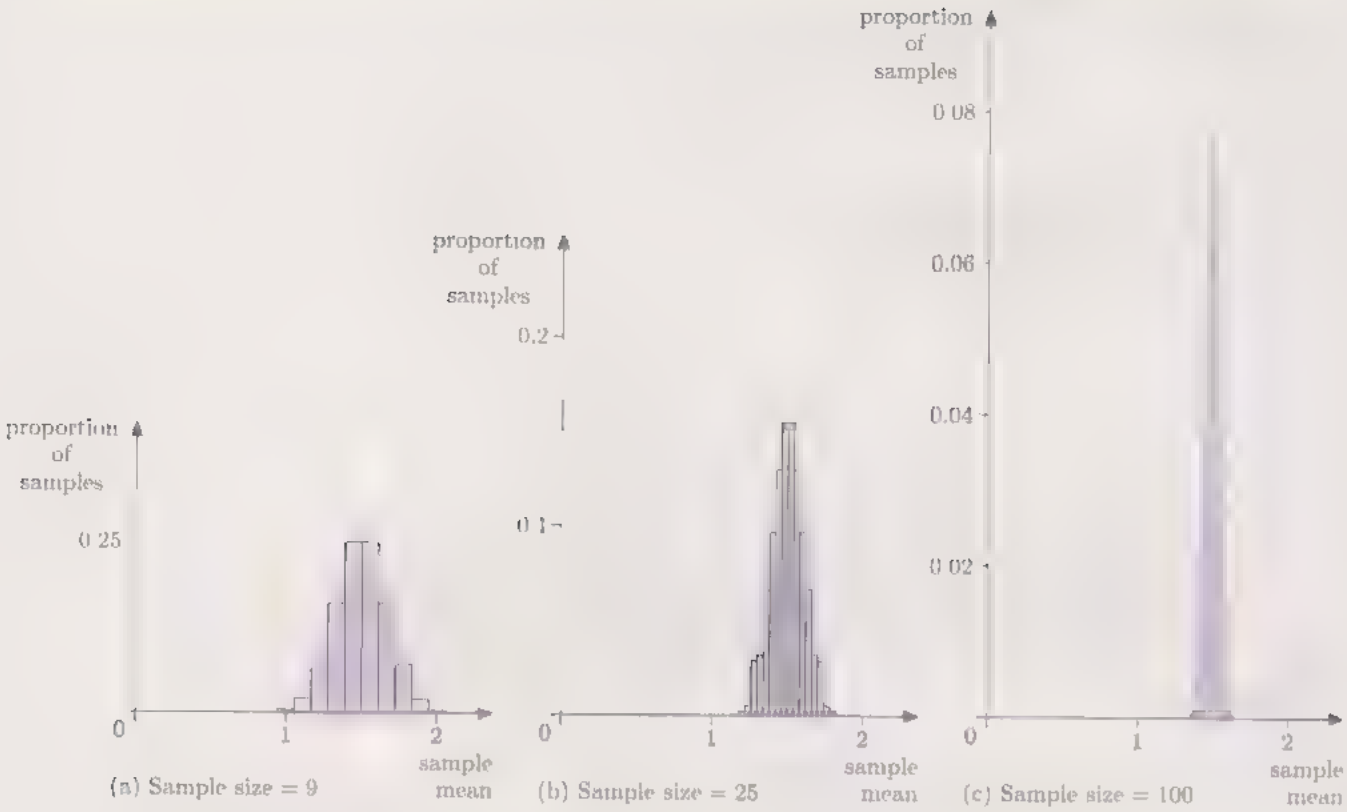


Figure 1.4 Sampling distributions of the mean

Activity 1.2 The shapes of sampling distributions

Describe the similarities among the sampling distributions in Figure 1.4. How does the shape of the sampling distribution of the mean change as the sample size increases?

Comment

All three sampling distributions are symmetrical about the value 1.5, and hence all have mean 1.5. In addition, they are all approximately bell-shaped. As the sample size increases, the sampling distribution rises more steeply to its peak at 1.5 and becomes more concentrated about this peak; that is, the spread of the sampling distribution decreases as the sample size increases.

The result that the spread of the sampling distribution of the mean decreases as the sample size increases is a useful one. It means that as you increase the sample size, the sample means vary less from the population mean; so the probability that the mean of a sample will be close to the population mean increases, and you can be more confident that the sample mean $\bar{x}$ is a good estimate of the population mean μ.

For instance, it is the case that the means of roughly 95% of samples of size 100 will be between 1.4 and 1.6 - quite close to the population

mean 1.5. (You will be able to check this for yourself later in this section in Activity 1.4.) For the motoring organisation, the more members it interviews, the more confident it can be that the results will be representative of the membership as a whole – the results are more likely to be accurate if a large number of members are interviewed. This is just what common sense would suggest: it is good to have our intuitions confirmed!

Of course, the results above are for the specific, and somewhat unrealistic, situation where half the members' households have one car and the other half have two cars. Would we obtain similar results for a different population distribution? Consider, for instance, the following situation.

Example 1.2 How many cars?

Suppose that the distribution of the number of cars per household for members of the motoring organisation is as shown in Figure 1.5. So some households have no car (for example, having sold their car since joining), a few have as many as four cars, but most have one, two or three cars.

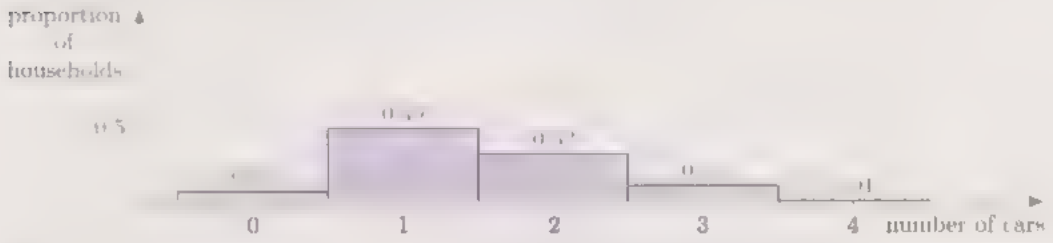


Figure 1.5 The population distribution

The mean number of cars per household for members of the motoring organisation is given by the mean of this distribution. The population mean μ is still 1.5 (as in the previous example). This distribution is not [illegible] for different sample sizes will look like in this case?

If you want to check this, [illegible]

Calculating the sampling distribution of the mean for samples of size 2 is more complicated for this population than it was for the population in the previous example, so we have not included all the details. To illustrate how proportions of samples with different means are calculated, we shall consider samples of size 2 with mean 0.5.

A sample mean of 0.5 arises only if the first household in the sample has no car (0) and the second household has 1 car, or vice versa. The probability that a household has no car is 0.07 and the probability that a household has 1 car is 0.49; so, using the rules of probability, the probability that the mean of a sample of size 2 is 0.5 is given by

$$0.07 \times 0.49 + 0.49 \times 0.07 = 0.0686.$$

Hence, the proportion of samples of size 2 with mean 0.5 is 0.0686. Other proportions are calculated in a similar way.

The sampling distribution is illustrated in Figure 1.6; its mean is 1.5. (This value was calculated using formula (4.4) from Chapter D1.)

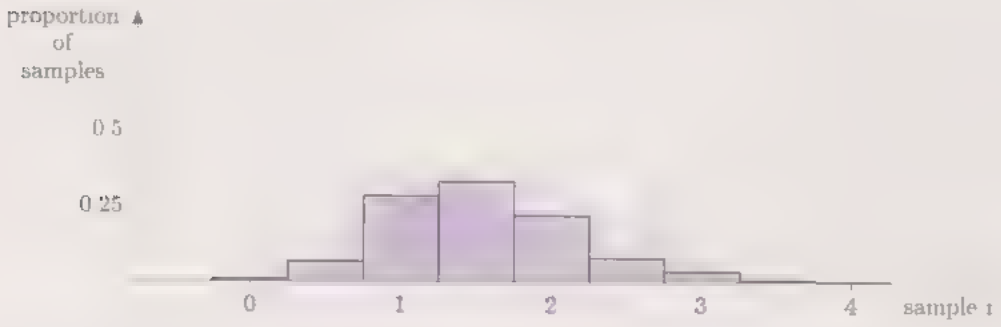


Figure 1.6 The sampling distribution of the mean for samples of size 2

The sampling distribution of the mean is not symmetrical, but it is less skewed than the population distribution. Moreover, its mean is equal to the population mean 1.5, and the distribution peaks at this value. It has a smaller spread than the population distribution.

The sampling distribution of the mean for samples of size 4 is shown in Figure 1.7; its mean is also 1.5.

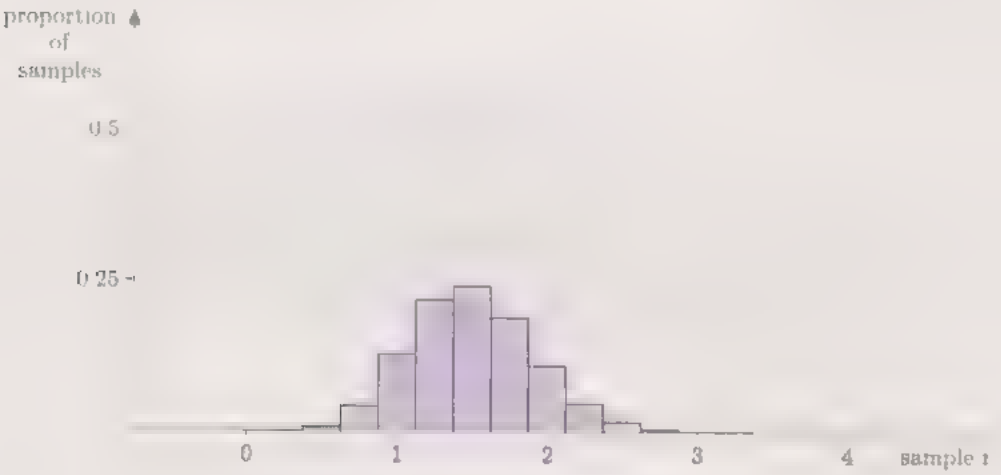


Figure 1.7 The sampling distribution of the mean for samples of size 4

Activity 1.3 The shapes of sampling distributions

Describe briefly the shape of the sampling distribution of the mean for samples of size 4 shown in Figure 1.7. Compare it with the population distribution (Figure 1.5) and with the sampling distribution of the mean for samples of size 2 (Figure 1.6).

Comment

The sampling distribution is roughly symmetrical with a peak at 1.5, the population mean. It is more peaked than both the population distribution and the sampling distribution of the mean for samples of size 2, and its spread is smaller.

There are two ways of calculating sampling distributions. In this subsection, they were found by using rules of probability - that is, using theory. However, the calculations can soon become very complicated and tedious! This is certainly the case for the second example, so, rather than using theory to find the sampling distributions of the mean for samples of sizes 9, 25 and 100 from this population distribution, a different method was used.

This second approach, which is used in the video band that you will be studying in the next subsection, is to simulate taking a large number of samples from a population. The sample mean is calculated for each sample, and the distribution of the sample means is observed. This is the empirical approach: in the long run, for each possible value of the sample mean, the observed proportion of samples with this sample mean should settle down to the theoretical proportion. The distributions of sample means obtained from simulations should be similar in shape to the theoretical sampling distributions, although because of sampling variation the empirical distributions are unlikely to be *exactly* the same as the theoretical distributions.

1.2 Sampling distributions

Your work in this subsection involves studying a video band about sampling distributions. The video looks at how the shape of the sampling distribution of the mean changes as the size of the samples increases. It also looks at how the spread of the sampling distribution of the mean, measured by its standard deviation, changes with sample size. In the video band, for each population distribution considered, the standard deviation of the sampling distribution of the mean is provided for various sample sizes, together with the standard deviation of the population distribution. You are not expected to be able to calculate these for yourself.

At various points in the video band, you will be asked a question about what you have just seen. You will then need to stop the tape while you write down your ideas. So *you will need to have pen and paper handy as you watch the video.* You may also find it useful to look at the graphics used in the video and to assess critically how useful they are for you in trying to get over particular ideas. When you come to produce your own images, you will then have a starting point for what you feel works in getting an idea across, and what is less useful.

Now watch Video Band D, 'Sampling distributions'.

Did you manage to find the relationship connecting the sample size and the standard deviation of the sampling distribution of the mean with the population standard deviation? For samples of size 100, the standard deviation of the sampling distribution is one tenth of the population standard deviation σ; for samples of size 4, it is one half of σ; for samples of size 25, it is one fifth of σ; and for samples of size 9, it is one third of σ. This suggests that the relationship involves the square root of the sample size: the standard deviation of the sampling distribution of the mean for samples of size n seems to be $\sigma/\sqrt{n}$, where σ is the population standard deviation.

If you identified this relationship, then you will have found that it works for samples of size 2 as well as 4, 9, 25 and 100.

In fact, the relationship holds for *all* sample sizes. The standard deviation of the sampling distribution of the mean (for samples of size n) is called

the **standard error of the mean** and is denoted by SE. So we can write the relationship succinctly as

$$SE = \frac{\sigma}{\sqrt{n}},$$

where σ is the population standard deviation and n is the size of the samples. Thus, if we know the standard deviation of the population, we can easily write down the standard deviation of the sampling distribution of the mean (for any sample size n). This, together with the other properties of sampling distributions discussed in the video, is summarised in the next subsection.

1.3 The Central Limit Theorem

In the video, for several different population distributions, you saw how the sampling distribution of the mean changes shape as the sample size increases. The main properties of sampling distributions discussed in the video may be summarised as follows.

- The mean of the sampling distribution of the mean is always equal to the population mean μ.
- The standard deviation of the sampling distribution of the mean (called the *standard error of the mean*, and denoted SE) is given by

  $$SE = \frac{\sigma}{\sqrt{n}},$$

  where σ is the population standard deviation and n is the size of the samples.

These results hold whatever the shape of the population distribution and whatever the size of the samples. They can both be proved mathematically, but the proofs are beyond the scope of this course.

- For large sample sizes, the sampling distribution of the mean may be approximated by a normal distribution. This approximation improves as the size of the samples increases.

In practice, the approximation is good for samples of size 25 or larger, whatever the shape of the population distribution. For some distributions, the approximation is a good one for smaller sample sizes. And if the population distribution is itself normal to begin with, then for any sample size the sampling distribution of the mean is also normal — no approximation is involved.

The three properties of sampling distributions given above together comprise the **Central Limit Theorem**.

For 'large' samples, say 25 or larger, the Central Limit Theorem can be used to give a range of values within which most sample means lie. This range can be calculated whatever the shape of the population distribution; the only information needed about the population distribution is its mean and standard deviation. The use of the Central Limit Theorem to find such a range of values is illustrated in the next example.

Example 1.3 The Central Limit Theorem at work

Suppose that samples of size 36 are drawn from a large population with mean 20 and standard deviation 3. By the Central Limit Theorem, the sampling distribution of the mean for samples of size 36 from this population is approximately a normal distribution with mean 20 and standard deviation given by

$$SE = \frac{3}{\sqrt{36}} = 0.5$$

In Section 6 of Chapter D2, you found that approximately 95% of values in a normal distribution lie within 1.96 standard deviations of the mean. So approximately 95% of sample means will lie within $1.96 \times SE = 1.96 \times 0.5$ of the mean 20, that is, between

$$20 - 1.96 \times 0.5 = 19.02$$

and

$$20 + 1.96 \times 0.5 = 20.98$$

So the means of approximately 95% of samples of size 36 from the population will lie between 19.02 and 20.98. Thus, roughly 95 times out of 100, the mean of a sample of size 36 taken from this population will differ from the population mean by less than 1.

Activity 1.4 The mean number of cars per household

In Example 1.1, it was assumed that the households of half of the members of a motoring organisation have one car and the other half have two cars. So the mean number of cars per household is 1.5. The standard deviation of the number of cars per household is 0.5 (you are not expected to be able to calculate this yourself). So $\mu = 1.5$ and $\sigma = 0.5$.

(a) Write down the mean and standard deviation of the sampling distribution of the mean for samples of size 100 from this population.

(b) Find a range of values within which the means of approximately 95% of samples of size 100 will lie.

Comment

The solution is given on page 37.

Activity 1.5 Sample means

The distribution of the heights of a population of men has mean 173 cm and standard deviation 7.2 cm.

(a) Write down the mean and standard deviation of the sampling distribution of the mean, for samples of 64 heights.

(b) Find a range of values within which the means of approximately 95% of samples of 64 heights will lie.

Comment

The solution is given on page 37.

Activity 1.6 More sample means

The distribution of the contents of packages labelled as containing 250 g has mean 252 g and standard deviation 5 g.

(a) Write down the mean and standard deviation of the sampling distribution of the mean, for samples of 40 packages.

(b) Find a range of values within which the mean weights of approximately 95% of samples of size 40 will lie.

Comment

The solution is given on page 37.

Activity 1.7 Two standard deviations

In this section, a result linking the standard deviations of two distributions has been discussed and used, namely

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population distribution and SE is the standard deviation of the sampling distribution of the mean for samples of size n drawn from the population. In the next section, we shall be discussing situations where a sample standard deviation is also used. So you need to keep clear in your mind what the various standard deviations are.

One way to distinguish between the different standard deviations is by using different notation for them, but of course you still need to remember what the notation represents. A table is sometimes a useful way of summarising information. So construct a table for the two different standard deviations involved in the formula above. Include in your table the terminology used for each standard deviation, what it is the standard deviation of, and its notation. You may also want to note down the result above which links the two standard deviations. You will be asked to add to your summary in the next section.

A Learning File Sheet is provided if you wish to use it.

Comment

A possible table is given on page 37.

Many techniques in statistics depend on making assumptions about the population distribution. The importance of the Central Limit Theorem lies in the fact that, when working with large samples, the sampling distribution of the mean is always approximately normal, whatever the population distribution. This means that, given a large sample, we can base our analysis on the normal distribution. The techniques described in the next section and in Chapter D4 depend on this result.

Summary of Section 1

In this section, you have seen how the means of samples from a population vary from sample to sample. The distribution of the means of samples of size n from a population is called *the sampling distribution of the mean for samples of size n*. In the video, simulations were used to investigate sampling distributions for samples of various sizes from a number of different distributions. You saw that the mean of a sampling distribution of the mean is always equal to the population mean μ, and the spread decreases as the sample size increases: the standard deviation of the sampling distribution of the mean, which is called the *standard error of the mean* (denoted by SE), is equal to $\sigma/\sqrt{n}$, where σ is the population standard deviation and n is the sample size. If the sample size is large, then, whatever the population distribution, the sampling distribution of the mean is approximately normal. These results together comprise the Central Limit Theorem, which is stated in the box below.

The Central Limit Theorem

For large sample sizes (at least 25), the sampling distribution of the mean for samples of size n from a population with mean μ and standard deviation σ may be approximated by a normal distribution with mean μ and standard deviation

$$SE = \frac{\sigma}{\sqrt{n}}.$$

In the next section, you will see how the Central Limit Theorem can be used to find a range of plausible values for the mean of a population, given only a sample of data. That is, you will see how to find a *confidence interval for a population mean*.

2 Estimating with confidence

In Chapter D2, a normal model was fitted to the data on the heights of 1000 Cambridge men. The sample mean $\bar{x}$ was used to estimate the population mean, and the sample standard deviation s was used to estimate the population standard deviation. In each case, a single number was used to estimate the value of an unknown parameter. The problem we shall address in this section is: *How good is the sample mean as an estimate of the population mean?* That is, by how much might the sample mean differ from the population mean? We shall approach this question by finding an *interval* of values within which we can be fairly confident that the population mean lies. This interval provides information about how inaccurate the sample mean might be as an estimate of the population mean.

We shall begin by using the sample of Cambridge men's heights to calculate a range of plausible values for the mean height of the population of all Cambridge men in 1902 — this is called a *confidence interval for the population mean*. To do this, we shall need the results for sampling distributions described in Subsection 1.3. As you will see, the method developed can be applied generally, given a large sample of data from a population, to calculate a confidence interval for the mean of the population.

Estimating the mean height of Cambridge men in 1902

Call the mean height in inches of all Cambridge men in 1902 — the population mean — μ, and the standard deviation of their heights σ. In Chapter D2, we gave the heights of a sample of 1000 men from this population.

Activity 2.1 The sampling distribution of the mean

(a) Write down the mean and standard deviation of the sampling distribution of the mean for samples of size 1000 from this population.

(b) Within what distance of the population mean μ will the means of approximately 95% of samples lie?

Comment

The sampling distribution of the mean for samples of size 1000 has mean μ and standard deviation $SE = \sigma/\sqrt{1000}$. So, since the sampling distribution of the mean is approximately normal, the means of approximately 95% of samples will be within $1.96 \times SE$ of the population mean; that is, they will differ from the population mean μ by less than

$$1.96 \times SE = 1.96 \times \frac{\sigma}{\sqrt{1000}}.$$

This is illustrated in Figure 2.1.

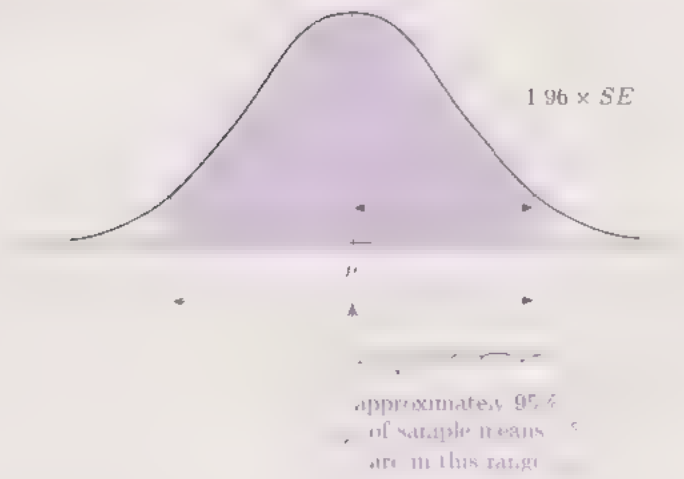


Figure 2.1 Sample means

So, for approximately 95% of samples of size 1000, the difference between the sample mean $\bar{x}$ and the population mean μ is less than $1.96\sigma/\sqrt{1000}$, or equivalently, for approximately 95% of samples of size 1000, the population mean will lie between the two values

$$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{1000}} \quad \text{and} \quad \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{1000}}.$$

The sample mean $\bar{x}$ can be calculated from the data. If we knew the value of σ, then we could calculate the values of the two expressions above and hence obtain an interval of values within which we could be fairly confident that the population mean μ lies. But the population standard deviation σ is unknown.

The way to overcome this difficulty is to estimate the value of σ. In Chapter D2, you saw that the sample standard deviation s can be used to estimate the population standard deviation σ; we shall use s to estimate σ here. Replacing σ by s leads to the following result.

Later in this section we shall discuss the effect of replacing σ by s.

For approximately 95% of samples of size 1000, the population mean μ lies between the two values

$$\bar{x} - 1.96 \times \frac{s}{\sqrt{1000}} \quad \text{and} \quad \bar{x} + 1.96 \times \frac{s}{\sqrt{1000}}.$$

Note that both the sample mean $\bar{x}$ and the sample standard deviation s vary from sample to sample. So different samples will give different values for these expressions.

The result above means that if we were to take a large number of samples of size 1000 from the population and, for each sample, calculate the interval of values

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{1000}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{1000}}\right), \tag{2.1}$$

then for approximately 95% of the samples we would find that the interval we had calculated did contain the population mean μ. For the other samples – approximately 5% of them – the intervals would not contain μ; they would 'miss' μ. An interval of values calculated using (2.1) is therefore called a **95% confidence interval for the population mean**.

For the sample of data in Chapter D2, the sample mean $\bar{x}$ is 68.9 inches and the sample standard deviation s is 2.57 inches, so a 95% confidence interval for the mean height of all Cambridge men in 1902 is given by

$$\left(68.9 - 1.96 \times \frac{2.57}{\sqrt{1000}},\ 68.9 + 1.96 \times \frac{2.57}{\sqrt{1000}}\right) = (68.7, 69.1),$$

rounding to one decimal place.

This interval is our answer to the question of how good the sample mean $\bar{x}$ is as an estimate of the population mean μ. In this case, the sample mean is 68.9 inches and this is our estimate of the mean height of all Cambridge men in 1902. But because of variation between samples, we know that this estimate may not be accurate. However, we can be 'fairly confident' that the true value of the population mean μ, the mean height of all Cambridge men in 1902, lies within the interval (68.7, 69.1).

Estimating the population mean

The method used to obtain a 95% confidence interval for the mean height of Cambridge men in 1902 can be used to find a range of plausible values for the mean of any population given a large sample of data from the population. Suppose that a population has mean μ and standard deviation σ, and that a sample of size n is taken from the population. By the Central Limit Theorem, if the sample size n is large (that is, 25 or more), the sampling distribution of the mean is approximately a normal distribution with mean μ and standard deviation $SE = \sigma/\sqrt{n}$. So the means of approximately 95% of samples of size n will be within $1.96 \times SE$ of the population mean μ; that is, the difference between $\bar{x}$ and μ will be less than $1.96 \times SE$. So, for 95% of samples of size n, the population mean μ will lie between

$$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}. \tag{2.2}$$

As before, we estimate the population standard deviation σ by the sample standard deviation s. This leads to the following result.

For approximately 95% of samples of size n, the population mean μ lies within the interval

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right). \tag{2.3}$$

This interval is called a *95% confidence interval for the population mean*. The end-points of the interval are called the **95% confidence limits**.

This use of the term 'limit' simply means 'bound', and is different from the sense of limit in 'Central Limit Theorem'.

You might have been wondering how good the sample standard deviation s is as an estimate for σ, and how replacing σ in (2.2) by s as in (2.3) affects the statement that the interval given by (2.3) is a *95%* confidence interval. You have seen that the sample mean varies from sample to sample less for large sample sizes than for small ones. In a similar way, it can be shown that the sample standard deviation also varies from sample to sample less for large sample sizes than for small ones - the larger the sample size, the better s is as an estimate of σ. Also notice that in (2.2), σ is divided by $\sqrt{n}$, so, for large sample sizes, replacing σ in (2.2) by s to obtain the confidence interval formula (2.3) is unlikely to introduce much inaccuracy. In addition, s will underestimate σ on some occasions and overestimate it on others, but overall, for *approximately 95%* of samples, the interval calculated using (2.3) will contain the population mean μ.

Activity 2.2 Sample size and interval width

(a) Two samples from a population are obtained one of size 50 and one of size 100, and each is used to calculate a confidence interval for the population mean. Which of the two intervals would you expect to be wider, and why?

(b) A sample of 100 components is taken from the output of a production line, and the resulting 95% confidence interval for the mean length of components is 2 mm wide. Roughly what size sample would be needed to obtain a 95% confidence interval which is only 1 mm wide?

Comment

(a) The width of a 95% confidence interval for a population mean is $2 \times 1.96 \times s/\sqrt{n}$, so it is directly proportional to $1/\sqrt{n}$. Hence you would expect the widths of 95% confidence intervals to decrease as the sample size increases. So a sample of size 50 would, in general, lead to a confidence interval wider than that for a sample of size 100 from the same population. Of course, the sample standard deviation s varies from sample to sample, so it is possible for a particular sample of size 100 to lead to a confidence interval wider than that for a particular sample of size 50; but, in general, a smaller sample will lead to a wider interval.

Since the width of a 95% confidence interval is directly proportional to $1/\sqrt{n}$, in general you would need to take samples four times as large in order to halve the widths of confidence intervals.

This result depends on the fact that, for any positive integer k,

$$\frac{1}{\sqrt{4k}} = \frac{1}{\sqrt{4}} \times \frac{1}{\sqrt{k}} = \frac{1}{2} \times \frac{1}{\sqrt{k}};$$

and hence, in general, confidence intervals calculated from samples of size $4k$ would be half as wide as those calculated from samples of size k.

So, if a sample of size 100 leads to a 95% confidence interval which is 2 mm wide, then a sample of 400 should lead to a 95% confidence interval which is about 1 mm wide.

The fact that, in general, a sample four times as large is required to halve the width of confidence intervals has practical implications. Collecting and analysing a large sample of data can be time-consuming and expensive, so sometimes a compromise has to be made between the desire for accuracy of estimation (a narrow confidence interval) and cost (which usually rises with sample size).

Before looking at some further examples, there are two points that we ought to mention.

First, you may have noticed that the formula for the calculation of a 95% confidence interval does not depend on the size of the population. The only assumption we have made about the size of the population is that it is much larger than the sample size. So the formula for a 95% confidence interval based on a sample of size 25 (say) is the same for samples taken from populations of size 10 000, 100 000 or 1 000 000. This means that you do not need to know the population size to calculate a confidence interval; you just need to know that it is much larger than the sample.

Second, the calculation of a confidence interval depends on the sampling distribution of the mean being approximately normal. As you saw claimed in Section 1, this is the case whatever the population distribution, provided that the sample size n is large. But what do we mean by 'large'? In practice, if n is at least 25, the approximation is quite good. So, in this case, 'large' means 'at least 25'.

The procedure for calculating a 95% confidence interval for a population mean given a large sample of data is summarised in the box below.

Given a random sample of size n from a population, a 95% confidence interval for the population mean μ is given by

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right),$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. The sample size n must be at least 25.

Example 2.1 The mean weight of male aquatic warblers

Source: Andrzej Dyrcz (1993) 'Biometrical differences between sexes in the breeding population of Aquatic Warblers Acrocephalus paludicola', *Ringing and Migration*, **14**, pp. 149–51.

A sample of 66 male aquatic warblers were caught and weighed in north-eastern Poland during their breeding season (in the early 1990s). The mean weight of these birds was 12.6 grams, and the sample standard deviation was 0.73 grams.

The mean weight of the population of all male aquatic warblers in the region is estimated to be 12.6 grams using the sample mean. But how much might this estimate differ from the population mean?

Since the sample size is large, the interval in (2.3) can be used to calculate a confidence interval for the mean weight of male aquatic warblers. A 95% confidence interval for the population mean weight in grams is given by

$$\left(12.6 - 1.96 \times \frac{0.73}{\sqrt{66}},\ 12.6 + 1.96 \times \frac{0.73}{\sqrt{66}}\right),$$

that is,

$$(12.4, 12.8) \quad \text{rounded to one decimal place.}$$

So we can be fairly confident that the mean weight of male aquatic warblers in north-eastern Poland is between 12.4 and 12.8 grams in the breeding season.

Activity 2.3 The mean weight of female aquatic warblers

(a) In the same study, 83 female aquatic warblers were caught and weighed. The mean weight of these birds was 12.1 grams, and the sample standard deviation was 0.87 grams. Estimate the mean weight of the population of female aquatic warblers, and calculate a 95% confidence interval for the population mean weight.

(b) Compare the confidence intervals for the mean weights of male and female aquatic warblers. What does this suggest about the weights of male and female aquatic warblers?

Comment

A solution is given on page 38.

Activity 2.4 Authorship and sentence length

Various criteria for establishing authorship have been investigated over the years. In 1940, C. B. Williams reported the results of an investigation he had carried out into sentence length as a criterion of literary style. He counted the number of words in each of roughly 600 sentences from each of three books, one by each of the three authors G. B. Shaw, H. G. Wells and G. K. Chesterton. The sentences were selected from various chapters and sections of the books.

Source: C. B. Williams (1940) 'A note on the statistical analysis of sentence length as a criterion of literary style', *Biometrika*, **31**, pp. 356–61.

The mean and standard deviation of the lengths of the 597 sentences written by G. K. Chesterton were $\bar{x} = 25.6$, $s = 10.76$. The sentences were from the book *A Short History of England* which was published in 1917. Calculate a 95% confidence interval for the mean sentence length in this book.

Comment

The solution is given on page 38.

Activity 2.5 Birth weights of babies

The weights of a sample of 28 full-term baby boys were measured to the nearest gram. The sample mean weight was 3490 grams, and the sample standard deviation was 452 grams. Calculate a 95% confidence interval for the mean weight of full-term baby boys.

Comment

The solution is given on page 38.

Activity 2.6 Radial velocities

Figure 1.5(c) of Chapter D2 shows the distribution of the radial velocities of 80 stars in a small region of the sky. The mean radial velocity was $-21\,\mathrm{km\,s^{-1}}$, and the standard deviation was $16.2\,\mathrm{km\,s^{-1}}$. Assuming that these stars may be regarded as a random sample of all the stars in this region, calculate a 95% confidence interval for the mean radial velocity of stars in this region of the sky.

Comment

The solution is given on page 38.

The 'average system' of weights and measures

Look along the shelves of any supermarket, or in your larder or refrigerator, and you will find numerous packets, tins and bottles with an 'e' mark on the label. The 'e' mark looks like the symbol in Figure 2.2, and is positioned close to the statement of the quantity contained in the package.

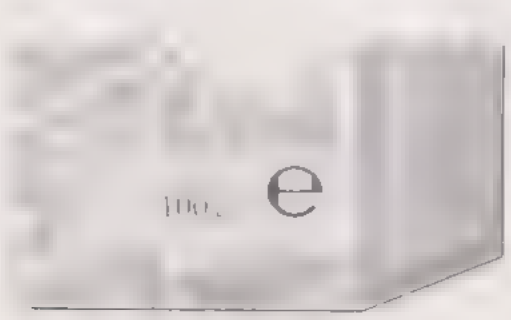


Figure 2.2 The 'e' mark

But what does the mark mean?

Essentially, the presence of the mark guarantees that a package has been packed according to the so-called 'average system' of weights and measures. This system ensures that the average (that is, the mean) contents of a group of packages is at least as great as the quantity stated on the label — the **nominal quantity**.

The average system is applied to most pre-packaged goods and is used throughout the European Union. Packages must be marked with the weight or volume of the contents, and must comply with three rules. The first rule says that the contents of packages must not, *on average*, be less than the nominal quantity. There is no guarantee that an individual package will contain at least this amount. However, the other two rules are designed to ensure that no package contains considerably less than the nominal quantity. (See the box opposite for further details of the average system. This information is included for interest only.)

In the UK, compliance with the rules is checked by Trading Standards Officers. But, in addition to this, packers check samples of packages themselves to identify any potential problems at an early stage.

Activity 2.7 Average weight

Punnets of strawberries are labelled as containing 250 grams on average. The packer weighs a sample of 30 punnets. The sample mean is 251.3 grams, and the sample standard deviation is 4.64 grams. Calculate a 95% confidence interval for the mean weight of all the punnets labelled as containing 250 grams. Can the packer be confident that the mean weight of all the punnets (not just the sample of 30 punnets) is at least 250 grams, as required by law?

Comment

The solution is given on page 38.

The average system of weights and measures

The average system controls the weight or measure of most pre-packaged goods. Packages must be marked with the weight or volume of the contents. Packers and importers must comply with three important rules.

1 The contents of packages must not, on average, be less than the amount marked on the label (known as the nominal quantity).

2 Not more than one package in forty may contain less than the nominal quantity by more than an amount called the *tolerable negative error* (TNE). This varies according to the nominal quantity; for example, 9 g on a 250 g pack. Such packages are called 'non-standard'. (This means that, for example, not more than one in forty of packages labelled as containing 250 g may contain less than 241 g.)

3 No packages may contain less than the nominal quantity by more than twice the TNE. These packages are called 'inadequate packages'.

The packer or importer must keep quantity control records, and must produce these for a local authority Trading Standards Officer on request.

A packer's main duty is to ensure that packages will pass a special statistical test called a 'Reference Test'. This is conducted by Trading Standards Officers, and shows whether the packer complies with the first two of the three rules. Packers must also meet the third of the rules, and either:

(a) make up all their packages on equipment listed in the regulations,

or

(b) check regular samples of the packages using equipment listed in the regulations, and keep a record of the checks for one year.

The packer has a duty to mark the packages legibly and permanently with the nominal quantity and the name and address of the packer or the person arranging for the packing to be done, or use a mark which will enable an inspector to identify the packer.

The 'e' mark must be at least 3 mm high and appear in the same field of vision as the statement of the nominal quantity. The mark constitutes a guarantee by the packer or importer that a package to which it is applied has been made up in accordance with the average system. There are restrictions on its use, and in most cases packers or importers who intend to export 'e' marked goods must notify their local weights and measures authority.

Source: This information was taken from a leaflet published by the National Metrological Unit. This unit ceased to exist in 1987.

There are two further questions that need addressing before we can consider this introduction to confidence intervals to be complete. First, what exactly does it mean to say that the level of a confidence interval is 95%? And second, what happens if the sample size is not large enough to use formula (2.3), that is, if $n < 25$? We shall consider these questions briefly now.

Confidence levels

Confidence intervals can be calculated for levels of confidence other than 95% — for example, 90% or 99%. These confidence intervals are just as straightforward to calculate as 95% confidence intervals. The formula for a 95% confidence interval for a population mean was obtained using the result that approximately 95% of values in a normal distribution lie within 1.96 standard deviations of the mean. To obtain a 90% confidence interval, we use the result that, for a normal distribution, approximately 90% of values lie within 1.64 standard deviations of the mean; this leads to the interval

See Section 6 of Chapter D2.

$$\left(\bar{x} - 1.64 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.64 \times \frac{s}{\sqrt{n}}\right).$$

Similarly, since approximately 99% of values in a normal distribution lie within 2.58 standard deviations of the mean, for a 99% confidence interval, 1.96 is replaced in formula (2.3) by 2.58 to give

$$\left(\bar{x} - 2.58 \times \frac{s}{\sqrt{n}},\ \bar{x} + 2.58 \times \frac{s}{\sqrt{n}}\right).$$

You will be asked to calculate only 95% confidence intervals in this course.

But what does it mean in practice to say that the confidence level of an interval is 95% or 90% or 99%? Consider, for instance, a 95% confidence interval. The fact that an interval calculated using formula (2.3) is a 95% confidence interval reflects the fact that on approximately 95% of occasions, that is, for approximately 95% of samples of size n, the procedure of calculating the interval

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$

will lead to an interval which contains the population mean μ. In Subsection 3.1, you will have the opportunity to verify this using the statistics software. You will be able to use the software to generate lots of random samples from a known population, and to calculate a 95% confidence interval for the population mean from each sample. You can then check that the proportion of 95% confidence intervals that do contain the population mean is approximately 95%.

90% and 99% confidence intervals are interpreted in a similar way. For instance, approximately 99% of intervals calculated using the formula for a 99% confidence interval for a population mean will contain the population mean; so only 1% of intervals will 'miss' the population mean. However, notice that increasing the confidence level produces a wider confidence interval (less precision). Similarly, reducing the confidence level leads to a narrower interval (greater precision).

What happens if n is small?

On the video, you saw that when the population distribution is normal, the sampling distribution of the mean looks approximately normal even for small sample sizes. In fact, the sampling distribution of the mean is always normal when the population distribution is normal. So, given a small sample of data ($n < 25$), can we use (2.3) to calculate a 95% confidence interval for the population mean when the population distribution is known to be normal? The answer to this is *no*. The reason for this is that, as already mentioned, the sample standard deviation s varies more from sample to sample for small samples than it does for large ones. So, for small samples, s is less reliable as an estimate of the population standard deviation σ; and hence results obtained using (2.3) for a small sample are unreliable, even when the underlying distribution is itself normal.

For small sample sizes, a different method is required for calculating a 95% confidence interval for the mean of a normal distribution. This was developed at the beginning of the 20th century; it is based on another family of continuous distributions, called t-distributions. Using this method, the formula obtained for a 95% confidence interval is very similar to the formula in (2.3): the only difference is that the number 1.96, which is a value obtained from the standard normal distribution, is replaced by the corresponding value from a t-distribution. (Provided that the population distribution is normal, this formula for a confidence interval may also be used when $n \geq 25$. However, for large values of n, this formula for a confidence interval and that in (2.3) lead to very similar results.) Confidence intervals based on t-distributions will not be discussed further in this course. But, if you study statistics further in the future, then you will almost certainly meet t-distributions and the corresponding confidence interval for the mean of a normal distribution. Different methods again exist for calculating confidence intervals using small samples from populations which are not normally distributed.

The interval in (2.3) can be used to find a 95% confidence interval for a population mean only when the sample size is at least 25. In this course, we shall be concerned with calculating confidence intervals only from 'large' samples of data, that is, for sample sizes of at least 25.

Activity 2.8 Another standard deviation

In this section, the Central Limit Theorem has been applied to produce a formula for a 95% confidence interval for a population mean. This formula involves the sample standard deviation s. Add the sample standard deviation to the table of standard deviations you began in Activity 1.7. If there are any results or notes that you wish to add to your summary, then do so now.

Comment

Some comments are given on page 38.

Activity 2.9 Same but different

As you read the summary of this section can you see any terms that fit the category 'Distinguishing different ideas with similar names'? If so, think how you could represent them clearly to a group of people who had little mathematical knowledge. Conversely, would your approach be different to more experienced mathematicians?

A Learning File Sheet is provided if you wish to use it.

Summary of Section 2

In this section, you have seen how a large sample of data from a population may be used to calculate a range of plausible values for the population mean – that is, a confidence interval for the population mean μ. A 95% confidence interval for a population mean is given by

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right),$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the size of the sample (which must be at least 25).

A 95% confidence interval may be interpreted as follows. If a large number of samples of size n are taken from a population, and the above formula is used to calculate a 95% confidence interval for each sample, then the population mean μ lies within approximately 95% of the intervals.

A narrower interval (greater precision) can be achieved either by increasing the sample size or by reducing the confidence level being used.

3 Confidence intervals on the computer

To study this section, you will need access to your computer and the statistics software.

In Section 2, a method was described for using a large sample of data from a population to calculate a range of plausible values for the population mean; this range of values is a 95% confidence interval for the population mean.

3.1 Interpreting a confidence interval

In this subsection, you will have the opportunity to use computer simulations to check the interpretation given in Section 2 of a 95% confidence interval.

Refer to Computer Book D for the work in this subsection.

3.2 Calculating a confidence interval

In this subsection, you will learn how to use OUStats to calculate a 95% confidence interval for a population mean from a large sample of data.

Refer to Computer Book D for the work in this subsection.

Summary of Section 3

In this section, you have used OUStats to calculate a 95% confidence interval for a population mean, given a large sample of data from the population. You have also used simulations to check the validity of the interpretation of a 95% confidence interval given in Section 2.

4 Why are opinion polls sometimes wrong?

At the time of the 1992 UK general election, none of the eve-of-election polls predicted that the Conservative party would win. The polls predicted that the Labour party would have a narrow lead over the Conservatives, but the Conservatives won the election by a comfortable margin. The predictions of the BBC's poll of polls and the election result, which were given in the introduction to this chapter, are repeated below.

	Conservative	Labour	Liberal Democrat
Poll of polls	38%	39%	19%
Election result	43%	35%	18%

The poll of polls quoted a 'margin of error' of ±1.5%. Roughly speaking, this means that, if the share of the votes received by any of the parties turned out to be as much as 1.5 percentage points more or less than the predicted share, this would not be regarded as surprising.

For example, the poll of polls predicted that the Conservative share would be 38%. So the margin of error indicated that a share of the votes as low as 36.5% or as high as 39.5% would not be surprising: any share in the range 36.5% to 39.5% was plausible. This range of plausible values is a confidence interval for the Conservative share of the votes. In fact, the polling organisations used a 95% confidence level for their calculations, so the range of plausible values, 36.5% to 39.5%, is a 95% confidence interval. Thus, the predicted share of the votes for a party and the margin of error can be used to write down a 95% confidence interval for the share of the votes that will be obtained by the party.

Activity 4.1 Confidence intervals

According to the poll of polls, what was the 95% confidence interval for the Labour share of the votes in the 1992 general election? What was the 95% confidence interval for the Liberal Democrat share of the votes?

Comment

The 95% confidence interval for the Labour share of the votes was (37.5%, 40.5%). And the 95% confidence interval for the Liberal Democrat share of the votes was (17.5%, 20.5%).

In the election, only the Liberal Democrat share of the votes was within the 95% confidence interval predicted by the poll of polls. The Labour and Conservative shares were both well outside their respective confidence intervals. Do you recall the interpretation of a 95% confidence interval for a population mean that was given in Sections 2 and 3? About 95% of samples from a population lead to 95% confidence intervals for the population mean which contain the population mean; 5% of samples lead to confidence intervals which do not contain the population mean, and so fail to predict the mean well.

A similar statement can be made for the predictions of opinion polls: for any particular party, 95% of polls can be expected to lead to predictions which differ from the actual result by less than the margin of error, and 5% of polls lead to predictions which differ from the actual result by more

than the margin of error. Roughly speaking, we can expect the pollsters to be wrong 5% of the time, even if there are no flaws in their methodology. This type of chance error, which is due to the variability which occurs between samples, is called **sampling error**.

So can the failure of the polls in 1992 be put down to sampling error, that is, to chance? Although in theory this is possible, in practice the Labour and Conservative shares of the votes were so far from their predicted shares that this is extremely unlikely to have been the case. For instance, at 43%, the Conservative share was 5% more than the 38% share predicted by the poll of polls, but the margin of error was only 1.5%. The probability of an error of this size occurring purely by chance is practically negligible. So what went wrong?

After the 1992 election, the Market Research Society of Great Britain established an enquiry into why the polls failed, and a report was published on what went wrong and some reasons were given. The enquiry identified three likely sources of error: unrepresentative samples, late swing and a phenomenon called *the spiral of silence*. These are each discussed very briefly below.

The samples used by the polling organisations are **quota samples**. In quota sampling, the population is divided up into categories, and the number of people interviewed in each category is supposed to reflect the proportion of the population in that category: for instance, three factors which are commonly used to divide the population into categories are age, gender and social class. A quota is set for the number of people to be interviewed in each category. It was found that in 1992 the information on which the quotas were based was inaccurate, and as a result the polls over-represented some groups and under-represented others. The samples were unrepresentative of those who voted on election day. It was thought that this could account for some, but by no means all, of the error in the polls.

The second factor thought to have had an effect was 'late swing'. A poll result is a snapshot of opinion at the time that the poll is carried out. If opinions change after this time, then the poll results will not forecast the election result accurately. In 1992, there seems to have been a swing from Labour to Conservative in the last few days before the election; this continued after the final interviews were conducted for the eve-of-election polls. Again, it was thought that this could account for a proportion of the error in the polls' predictions.

However, these two sources of error did not fully explain the size of the error that occurred. A third factor is thought to have been a phenomenon known as *the spiral of silence*. Roughly speaking, this can be described as follows: some people who intend to vote for a party which is perceived to be unpopular are reluctant to admit this. So a greater proportion of the people intending to vote for the 'unpopular' party than of those intending to vote for other parties fail to say so. This can introduce bias into the results of the opinion polls.

All three sources of error are described in greater detail in the television programme *The Spiral of Silence*. The programme also discusses how the polling organisations are modifying their procedures in an attempt to eliminate these sources of error from future polls.

Summary of Section 4

In this brief section, various factors which may have contributed to the failure of the polls to predict accurately the result of the 1992 general election have been discussed briefly. These include the unrepresentativeness of the samples, late swing, the spiral of silence and sampling error (that is, chance). You have seen how a poll's predictions can be used, together with its margin of error, to give a 95% confidence interval for a political party's share of the votes.

If you are interested in learning more about political opinion polls, then you will find a readable account of their history and methodology in a paperback by Robert M. Worcester called *British Public Opinion: A Guide to the History and Methodology of Political Opinion Polling.* This is published by Basil Blackwell in the series 'Making Contemporary Britain'.

Summary of Chapter D3

In this chapter, the idea of a sampling distribution has been introduced. Part of your work involved studying a video band in which simulations were used to obtain the sampling distribution of the mean for samples of various sizes and from several populations; you were invited to conjecture general results about sampling distributions from these special cases. The properties of sampling distributions illustrated in the video band are consequences of the Central Limit Theorem. This important theorem enables us to make inferences about a population from a large sample of data from the population.

This chapter has been concerned with estimating estimating a political party's share of the vote in an election and estimating an unknown population mean. You have seen how the Central Limit Theorem can be used to obtain a range of plausible values for an unknown population mean - a confidence interval for the population mean. You have also seen the connection between the 'margin of error' quoted in an opinion poll and a confidence interval. In the next chapter, you will see how the Central Limit Theorem can be used to provide a way of testing whether there is a difference between two population means.

Before you move on to the next chapter, it is important for you to make sure that you have understood and assimilated the main ideas introduced in this chapter. To help you reflect on these ideas and on what you have learned, try the activity suggested below.

Activity 5.1 Statistical themes

The chapter has introduced a number of statistical ideas. Write down in your own words what you consider to be the most important statistical idea that you have encountered while reading this chapter. What did you find most interesting? What did you find most difficult?

A Learning File Sheet is provided if you wish to use it.

Comment

Some comments are given on page 38.

Learning outcomes

You have been working towards the following learning outcomes.

Terms to know and use

Sampling distribution of the mean, standard error of the mean, confidence interval for a population mean, margin of error, sampling error, quota sample, confidence level.

Symbols to know and use

The notation SE for the standard error of the mean.

Ideas to be aware of

- The way the shape and the standard deviation of the sampling distribution of the mean change as the size of samples is increased.
- The Central Limit Theorem.

Features of the statistics software to use

- Run simulations of samples from a known distribution to obtain confidence intervals.
- Use OUStats to obtain a 95% confidence interval for the population mean given a large sample of data from a population.

Investigating processes to aid understanding

- Conjecture a general result from results obtained in special cases.

Solutions to Activities

Solution 1.1

The completed table is shown below.

Sample values	Sample mean	Proportion of samples
1 1 1 1	1	$\frac{1}{16}$
1 1 1 2	1.25	$\frac{1}{16}$
1 1 2 1	1.25	$\frac{1}{16}$
1 1 2 2	1.5	$\frac{1}{16}$
1 2 1 1	1.25	$\frac{1}{16}$
1 2 1 2	1.5	$\frac{1}{16}$
1 2 2 1	1.5	$\frac{1}{16}$
1 2 2 2	1.75	$\frac{1}{16}$
2 1 1 1	1.25	$\frac{1}{16}$
2 1 1 2	1.5	$\frac{1}{16}$
2 1 2 1	1.5	$\frac{1}{16}$
2 1 2 2	1.75	$\frac{1}{16}$
2 2 1 1	1.5	$\frac{1}{16}$
2 2 1 2	1.75	$\frac{1}{16}$
2 2 2 1	1.75	$\frac{1}{16}$
2 2 2 2	2	$\frac{1}{16}$

Solution 1.4

(a) The mean of the sampling distribution of the mean for samples of size 100 is equal to the population mean: 1.5. The standard deviation is given by

$$SE = \frac{0.5}{\sqrt{100}} = 0.05.$$

(b) Approximately 95% of sample means will lie between

$$1.5 - 1.96 \times 0.05 = 1.402$$

and

$$1.5 + 1.96 \times 0.05 = 1.598.$$

This agrees with the statement made after Activity 1.2 that the means of roughly 95% of samples of size 100 from the population will be between 1.4 and 1.6.

Solution 1.5

(a) The mean of the sampling distribution of the mean for samples of size 64 is equal to the population mean: 173 cm. The standard deviation is given by

$$SE = \frac{7.2}{\sqrt{64}} \simeq 0.9\,\text{cm}.$$

(b) Approximately 95% of sample means will lie within $1.96 \times SE$ of the population mean, that is, between

$$173 - 1.96 \times 0.9 \simeq 171.24\,\text{cm}$$

and

$$173 + 1.96 \times 0.9 \simeq 174.76\,\text{cm}.$$

Solution 1.6

(a) The sampling distribution of the mean for samples of size 40 has mean 252 g and standard deviation

$$SE = \frac{5}{\sqrt{40}} \simeq 0.79\,\text{g}.$$

(b) Approximately 95% of sample means will lie between

$$252 - 1.96 \times 0.79 \simeq 250.45\,\text{g}.$$

and

$$252 + 1.96 \times 0.79 \simeq 253.55\,\text{g}.$$

Solution 1.7

A possible structure for a summary table is shown below.

Terminology	Notation	Standard deviation of ...	Useful results
Population standard deviation	σ	population	
Standard error of the mean	SE	sampling distribution of the mean	$SE = \frac{\sigma}{\sqrt{n}}$, where n is the size of the samples

Solution 2.3

(a) The sample mean was 12.1 grams, so this is an estimate of the mean weight of the population of female aquatic warblers. A 95% confidence interval for this mean weight in grams is

$$\left(12.1 - 1.96 \times \frac{0.87}{\sqrt{83}},\ 12.1 + 1.96 \times \frac{0.87}{\sqrt{83}}\right),$$

that is, (11.9, 12.3) rounded to one decimal place.

(b) The interesting point about the confidence intervals for the mean weights of male and female aquatic warblers is that they do not overlap. The values in the confidence interval for the males are all higher than *any* of the values in the confidence interval for the females: (12.4, 12.8) for the males compared with (11.9, 12.3) for the females. This suggests that males are heavier, on average, than females. In the next chapter, a test is described which will enable you to investigate whether the difference between the mean weights of the samples of males and females could be due to sampling variation, or whether it is likely that there really is a difference between the mean weights of the populations of male and female aquatic warblers.

Solution 2.4

A 95% confidence interval for the mean length of sentences in G. K. Chesterton's book is given by

$$\left(25.6 - 1.96 \times \frac{10.76}{\sqrt{597}},\ 25.6 + 1.96 \times \frac{10.76}{\sqrt{597}}\right),$$

that is, (24.7, 26.5) rounded to one decimal place.

Notice that, even though the lengths of sentences are very variable ($s = 10.76$), the confidence interval is short, so that we can be fairly sure that 25.6 is a good estimate of the mean sentence length. The confidence interval is short because it was based on a large sample of data ($n = 597$).

In Subsection 3.2, you will have the opportunity to investigate the data on sentence lengths for all three authors using OUStats.

Solution 2.5

A 95% confidence interval for the mean weight of full-term baby boys is given by

$$\left(3490 - 1.96 \times \frac{452}{\sqrt{28}},\ 3490 + 1.96 \times \frac{452}{\sqrt{28}}\right),$$

that is, (3323, 3657) rounded to the nearest gram.

Solution 2.6

A 95% confidence interval for the mean radial velocity in $\mathrm{km\,s^{-1}}$ of stars in this region is given by

$$\left(-21 - 1.96 \times \frac{16.2}{\sqrt{80}},\ -21 + 1.96 \times \frac{16.2}{\sqrt{80}}\right),$$

that is, (−24.5, −17.5) rounded to one decimal place.

Solution 2.7

A 95% confidence interval for the mean weight of all the punnets packed is given by

$$\left(251.3 - 1.96 \times \frac{4.64}{\sqrt{30}},\ 251.3 + 1.96 \times \frac{4.64}{\sqrt{30}}\right),$$

that is, (249.6, 253.0) rounded to one decimal place.

Since the confidence interval extends below 250 grams, it is quite plausible that the mean weight of all the punnets is less than 250 grams. Perhaps the packer should weigh a few more punnets to investigate this possibility.

Solution 2.8

Something like the following could be added to the table begun in Activity 1.7.

Terminology	Notation	Standard deviation of ...	Useful results
Sample standard deviation	s	sample	For $n \geq 25$, a 95% confidence interval for a population mean is $\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$.

Solution 5.1

The most important idea discussed in this chapter is the Central Limit Theorem. The idea of a confidence interval is also important. However, the procedure for calculating a confidence interval for the mean of a population given a large sample from the population is obtained by using the Central Limit Theorem. This is just one out of a large number of applications of the Central Limit Theorem. The technique discussed in the next chapter is another example of an application of the theorem.

Using Mathematics

The Open University
ISBN 0 7492 7890 0